Ethical Online AI Systems through Conscientious Design

Pablo Noriega
IIIA-CSIC

Harko Verhagen
Stockholm University

Julian Padget
University of Bath

Mark d'Inverno
Goldsmiths, University of London

***Abstract*—There is increasing interplay between humans and artificial intelligent (AI) entities in online environments. With the growing autonomy and sophistication of these AI systems, the hybrid communities which are formed start to behave like the more-familiar, human-only social systems. This sets up the challenge to find systematic ways to ensure reliable governance for these interactions just as we do in human communities. This article proposes a novel approach to building governance for hybrid communities using what we call *Conscientious Design* (CD). There are two key aspects to CD: (i) the introduction of *value categories* that guide the identification of relevant stakeholder values, coupled with (ii) a tripartite model for online institutions that serves to describe the interactions of hybrid communities of humans and artificial (AI) entities in a way that is consistent with the values of all stakeholders.**

Introduction

We are entering a time of increasing interaction with artificial intelligent systems (AIS) in online environments. Moreover, these interactions will be increasingly complex as we become more familiar with inhabiting such communities and the artificial systems themselves become more sophisticated and autonomous. This suggests a pressing need to explore approaches to the design of such systems to build confidence that the emerging online environments and behaviours are places we would wish to inhabit.

The greater autonomy of artificial entities means increased potential for them to influence the social and psychological states of human participants. This growing potential raises new concerns about how one can protect participants' well-being when these more complex computational agents can be more incompetent, untrustworthy and – even – malevolent.

Whilst engineering ethical considerations into AIS is often spoken about, current practice is patchy at best. Even if ethics are considered, there are no systematic or principled means to ensure that the design and implementation of a system with convincing answers to questions such as: what does it mean to do the "right" thing?, how can it be known with any degree of certainty that a new AIS will support the "right" thing?, and when is enough "enough" in terms of what needs to be thought about?

Furthermore, the risks of getting it "wrong", and a new system causing harm, are hard to assess too. Not only because all kinds of unplanned

 Published by the IEEE Computer Society

behaviours and impacts could emerge, but also because of a lack of documented experience in addressing ethical concerns in AIS design. Because considering these factors together is so hard, it might lead to ignoring the issue altogether or hoping that basic common sense will be enough to resolve any problems on the fly.

In response to these concerns, we have developed the notion of *Conscientious Design (CD)* as a systematic and practical approach to support practitioners in the ethical design of AIS. It is an approach that builds on well-established practices in value-sensitive design (VSD) [1], Alexander's "habitable spaces" [2], and Deming's total quality management (TQM) [3]. It also provides a way of using familiar agile concepts to imbue values in AISs. Additionally, it puts human and artificial agent participants in control of co-evolution of the online spaces they jointly inhabit.

Participants in Alexander's habitable spaces are physically constrained whereas in online systems they are constrained in different ways. First, they are constrained by the platform itself and what actions it allows, known as platform-provided affordances [4] (e.g., "buy", "like", "ban"). Actions not provided by the platform simply cannot take place. Second, actions of one participant are constrained by the normative expectations that the other participants have of what is acceptable or unacceptable behaviour (e.g., spamming, helping, ignoring), where non-compliance may lead to sanctions against the acting agent. These two categories of constraint are perhaps most easily understood through our own experiences of using on-line platforms (e.g., shopping, social networks).

We base our proposal for CD on a particular subclass of AIS that we have been researching for some years, called *online institutions (OIs)* [5], [6]. OIs contain policies that facilitate the governance of participant activity, either through what a participant is allowed to do in certain circumstances or what a participant may choose to do or not to do for the sake of any social consequences. Online institutions embody both affordances and norms, interpreting Alexander's "Timeless way of building" for the social – often commercial – spaces in which we participate on the Internet. Furthermore, OIs (as with all AISs) are software constructs, and so have an intrinsic adaptability and resilience, which means that they can in theory support Deming's evolutionary approach to the achievement of quality over time, founded on VSD's value principles. We also take the position that by considering online institutions we can most effectively map out the principles and building blocks of conscientious design, which can then be applied to a wider class of AIS in due course.

Conscientious Design

Stakeholders in VSD are presented with a simple ethical framework: first consider what is right, and secondly what is good [1], which hints at a hierarchy of values and debates over which values are right and which are good. This creates two challenges: how to identify the (small) set of core values and to which value or values to associate different aspects of the design, without connecting everything to everything.

CD builds on VSD by providing a value framework from within which to argue about the "how and why" of stakeholder values, rather than whether one value is more important than another. The framework involves three *value categories* (thoroughness, mindfulness, and responsibility), a *systematic identification of contexts* (through the *WIT pattern*) where these categories are instantiated in OIs, and a process to make *values operational*.

The CD value categories are:

- **Thoroughness**: this refers to conventional technological values that promote the technical quality of the system. In any (standalone) system, values include completeness and correctness of the specification and implementation, reliability and efficiency of the run-time version of the system, robustness, resilience, accessibility, and security. Thoroughness also applies (in the "situated OI") to the technological compatibility of the OI with the context where it is embedded, as well as its integrity (intrusions and data or communication corruption);
- **Mindfulness**: We have chosen this word carefully to respond to the considerations about impact on human users that are so often over-looked. In its characterisation mindfulness includes building a wider awareness of what is happening to humans and society

		Thoroughness	Mindfulness	Responsibility
EU HLEG Guidelines for Trustworthy AI Ethical Principles	**Human autonomy**	*respect of user's preferences* flexible service options unobtrusive interface	*respect of user's attention* quick trip negotiation	*respect of user's freedom* fair contract termination clauses
	Prevention of harm	*risk minimisation* risk / liability driver's qualifications	*certainty of commitments* assurance of identities of users before trip	*compatibility with local culture* reliable and convenient payment options
	Fairness	*user neutrality* uniform role-based criteria	*fair service allocation* best car best time	*fair fares* fair and competitive fares
	Explicability	*thorough representation* enough indicators for ongoing value assessment	*reliable and accessible use relevant information* pricing algorithm should explain itself	*legal and fiscal responsibility* clear invoices
General Principles of Ethically Aligned Design for Autonomous / Intelligent Systems	**Human rights**	*bias free* good quality system-generated data	*protection of users' rights* drivers get paid	*environmentally responsible* vehicle emissions requirements
	Well-being	*integrity of user identity* user access and service completion validation	*user satisfaction* driver compensations and rewards	*passenger and driver safety* insurance
	Data agency	*integrity of system data* removal of app implies removal of server-side data	*clear and enforceable data use policies* user's choice over user-associated data ownership	*fair third-party data access* foreclosure diagnostics
	Effectiveness	*solid payment methods* credit verification/application	*quick negotiation* certainty about mutual commitments	*liability protection* insurance
	Transparency	*adequate value assessment indicators* elicit WIT-contexts stakeholders' values	provide relevant and accurate information keep user- relevant history	*compatibility with social norms* data ownership policies
	Accountability	*reliable value attainment assessment* indicators for historic and ongoing assessment	*availability of user-relevant and accessible information* passenger's / driver's activity history	*compliance with local regulations* auditable transactions
	Awareness of Misuse	*reliable interfaces* thorough input/output checkpoints	*protection of users needs* user-satisfaction elicitation	*awareness of potential illegal activity* collusion among car-owners
	Competence	*objective performance assessment* correct app updating	*alignment with business objectives* cost/benefit analysis cash flow	*technological compatibility* accurate mapping of available cars

Table 1: Mapping EU [7] and IEEE [8] principles onto the three CD value categories Thoroughness, Mindfulness, and Responsibility.
Italics denote examples of operationalized values and plain text the indicators of these values.

through the use of technology to guide us in making the right choices, in line with Deming's principles. Examples of values in this category concern data ownership (privacy, data agency, usage traces), and well-being (accessibility, respect of user's attention);

- **Responsibility**: these are values that address the effects towards the owner, the users, and external stakeholders (regulators, suppliers, partners,...) of using the OI. Here, we can also include the effects of the system on the context in which it is situated (liability, accountability), and how that context may affect intended users, designers and owners (legitimacy, user protection, no hidden agency).

In broad terms, CD's contributions are the distinctive attention to policies that govern interactions the systematic separation of analysis by stakeholder, context and time supported with the WIT pattern. In particular, the CD proposal supports the initiatives from the EU [7] and IEEE [8] on building AIS. Indeed, these initiatives underline the timeliness of CD. In Table 1 we illustrate how CD values relate to the EU and IEEE principles respectively, based on the keywords used in the documents in which they are described. For instance, the EU Guidelines have under the ethical principle of explicability the following example measures: "traceability, auditability and transparent communication on system capabilities" [7]. These belong to the CD value of responsibility, in that they describe the anchoring of the system. As an example of mapping IEEE ethical design principles, consider competence. This addresses safe and effective operation [8], i.e., it belongs to the CD value of thoroughness, with its focus on the technical quality of the system.

Apart from showing how to map all EU [7] and IEEE [8] principles onto the CD proposal, Table 1 also shows that these principles can be mapped onto all CD value categories. Thus, CD value categories support more than one way of looking at each particular principle. This is a notable benefit of CD's principled approach.

Online institutions and the WIT pattern

CD aims to help designers in debating the why and what of the system, but the translation from "what" to "how" needs equally careful handling to maintain the separations of concerns

suggested above. As with Alexander's blueprints, the objective is not to provide an answer, but a way to think about the answer and arrive at an appropriate solution every time. Therefore, we propose the World-Institution-Technology design pattern for OIs (see Fig. 1), where the world (W) is a social space that is a sub-context of the real world, institutions (I) are the policy frameworks into which the values that characterise the system are imbued, and the technological space (T) where online interactions are processed according to software representations of the institutional conventions.

Online institutions in CD are the glue that binds W, I and T together, to mirror the functions of conventional social and economic institutions [9]. This subclass of sociotechnical systems is formally defined in [10], [11] and is a refinement of other abstractions of systems for social coordination and artificial or electronic institutions [12], [5]. Informally, an OI provides technological support for human and software agents to interact online with each other, and establishes the policy – the "rules of the game" – that governs those interactions. The terms of the policy determine what fragment of the real world is relevant, what events and actions that take place in the world are recognised by the institution and what their effects in the institution are, and vice versa. For this purpose, an OI (i) maintains an institutional state that is accessible to all the active participants and (ii) may recognise whether an action is correct (in the prevailing circumstances) and, if so, update the institutional state accordingly. For example, if a customer signals – via the *Uber* app – that a pick-up is not taking place, the system would ignore the signal if a driver is about to arrive or notify the customer that another driver is on its way.

We now look in more detail at the relationships between the components of the WIT pattern (Fig. 1):

- W↔I: intuitively, I is an abstraction of the relevant sub-context in W, that captures "just enough" of the real-world dynamics – the actions and events that can occur that matter for the sub-context, like movement, or picking up or dropping items in a game – and an institutional model that represents the policy that applies to those recognised actions and events. In the other direction, institutional changes need grounding to have consequences in the social (world) context, such as a passenger rating affecting driver selection in *Uber*;
- I↔T: the abstraction in I provides the specification for what must happen in T, telling the developer what function the technology space should deliver, while the relationship in the other direction documents how the technology space implements what I specifies;
- W↔T: the relationship between W and T that enables the participants of the social (world) context to interact, by whatever interfaces are appropriate (webpages, phones, game handsets, VR, sensors of various kinds) providing inputs to the OIs (actions and events) and receiving outputs (institutional interpretations and consequences of those actions and events).

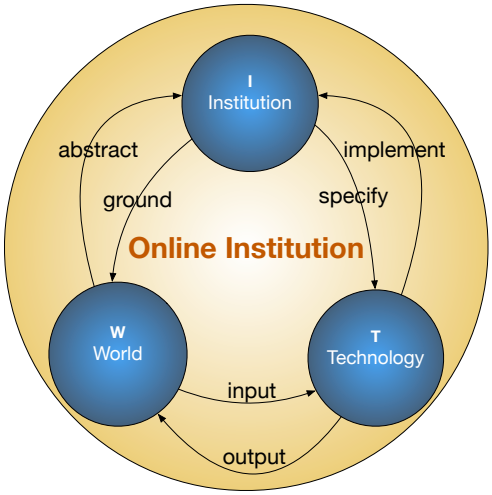

Figure 1: The WIT Tripartite Pattern: the World, Institution, and Technology Views and the relationships between them (after [13]).

Moreover, the WIT pattern also helps to differentiate what to consider when examining the legal, social, and technological compatibility of the OI as a system that is situated in its (evolving) working environment [11].

Putting values into play

Values are powerful and practical devices to imbue ethical behaviour in AIS. In general, values help assess the "worthiness" of a state of affairs and to decide the "right" action to take [14]. Institutional governance should promote or require actions whose effects align with stakeholders' values and prevent or discourage those that do not. We propose a three stage process for making values operational:

1. Value interpretation consists of identifying behaviours and outcomes that are characteristic of that value so that these are encouraged or guaranteed to happen. The three CD value categories must be instantiated with concrete values that allow for a refinement of its interpretation, implementation, and assessment. Interpretations of the same value may vary depending on the context in which the behaviour is to be observed, the perspective of the stakeholder who observes it, and the moment when the value is assessed.

The WIT pattern facilitates this analysis with the identification of the different contexts. There are two approaches for defining the meaning of a value. One, is to produce an explicit description of behaviours that uphold (or demote) the value, the other is to choose a set of indicators – observable parameters in the state of the system – that reflect support for the value (or its demotion).

2. Value implementation can be achieved by focusing on the behaviours and outcomes aligned with the value. There are three typical strategies of implementing values. They are not mutually exclusive and strategy selection is a design decision. The three strategies are:

- *Hard-wire constraints* and procedures that implement specific behaviour and indicators associated with the interpretation of values. This presupposes the choice of the relevant entities that provide the basis for the institutional model and its implementation. This hard-wiring needs to adapt to the evolution of an OI. For instance, in online multiplayer games such as League of Legends, the base capacities and skills of the characters the players can choose from are given, as are the ways in which these can be extended during game-play;
- *Use explicit policies* (e.g. technical norms [15]. These may comprise (i) functional norms that specify the preconditions and the effects of admissible actions; and can be easily linked with indicators; or (ii) procedural norms that define how to perform and implement a specific behaviour that interprets a value For example, in *Uber*, a "fairness" norm assigns a rider the closest available car but prioritises cars with higher client satisfaction ratings;
- *Influence the decision-models of participants* by providing additional information or arguments that may promote a change of decisions. In online games, such as League of Legends, the problem of toxic gaming and inappropriate language between temporary teammates is detrimental to enjoyment. In League of Legends, at first a sanctioning strategy was chosen – initially using selected human players as a jury to judge complaints [16], later replaced by an automated sanctioning system which was criticised, amongst other reasons, for not being transparent [17]. In its latest incarnation, a positive reward system has been put in place as an honour system in which team mates can give each other positive feedback. How this feedback is represented in the game (a badge with a numerical value) and what it may result in (extra in-game rewards) has changed over time but an overall critique remains to this system as well: it is the game company who decides what is and what is not transgressing the "honour rules of the game" [17], i.e., not all stakeholders are part of the discussion on how to assess the fulfilment of the value of "fun";

3. Value assessment determines to what degree a value is being attained. This may either be validating that a required behaviour is achieved, or measuring value indicators. Since value interpretation (and implementation) is "context dependent" we put all the needed assessment components into a *Value Assessment Framework* that, for each stakeholder, consists of: (i) the values that are relevant in the specific assessment context; (ii) the corresponding interpretation and validation/measuring mechanisms for each value;

and (iii) the aggregation function for the set of values.

Concluding Remarks

We are all aware of increasing interaction in online communities of human and software participants. Many of these have been designed and implemented without truly recognising that a new kind of responsibility in software design is needed to protect human well-being.

In this article we have outlined conscientious design (CD) as our response to this need. We have specifically applied CD to *online institutions* which are a subclass of AISs, where governance is explicitly represented and enforceable.

Our intention in proposing CD is to support developers of ethical hybrid online social systems in three ways.

First, to provide a blueprint for the construction of online systems that we would be happy to inhabit. This blueprint is achieved through the separation of world, institution, and technological concerns using the WIT pattern to facilitate the design of online institutions.

Second, we propose three value categories – thoroughness, mindfulness and responsibility – and provide a characterisation and justification for each. These provide a high-level guide to embedding the shared and agreed values of stakeholders in OIs. We believe that any consideration of ethical issues should consider these three categories in detail.

Third, to enable the design of explicit, transparent governance mechanisms that contain mutually comprehensible representations of human authored policies to say what participants may do under what circumstances.

We believe these considerations together support the explicit consideration of ethical aspects. They enable stakeholders, including designers, to explicitly introduce their own values into the design of ethical AISs. It enables a balanced focus on the affordances and norms that are so critical in understanding governance. CD enables the system to adapt transparently as the needs and value priorities of stakeholders change over time.

In closing we set out why the CD approach matters:

- CD is *principled*: It provides an intuitive way to operationalize the principles set out in the trustworthy AI [7] and ethically aligned design [8] guidelines;
- CD *reorients existing methods* for AIS. CD extracts elements from value-sensitive design, design patterns, and process quality to apply known thinking from agile development to target a class of internet-based systems;
- CD is *timely* because we are in the early stages of the construction of online sociotechnical systems that have both human and AIS participants;
- CD is *practical*: value imbuing is not a trivial process but our experience shows that it can be tackled with a principled strategy that interprets conscientious values in relevant contexts (stakeholders, stand-alone, situated) and uses adequate devices for making them operational (value interpretation, instrumentation, measurement, aggregation);
- CD is *malleable*. It requires an ongoing implementation process involving stakeholders from the start. Values are not set in stone; with CD, they are identified and fit (ex-ante) to the specific context and are progressively assessed and adapted ex-post;
- CD facilitates *continuous improvement* – as modifications or add-ons – for refactoring conscientious values into existing systems.

We hope this work can be the start of building an interdisciplinary community of researchers and practitioners who can join forces to further develop the body of CD practice with rigorous descriptions of (re-usable) CD components, documenting use cases that embed ethical considerations in the design process, and so building better, fairer, and safer online worlds.

REFERENCES

1. B. Friedman, D. G. Hendry, and A. Borning, “A survey of value sensitive design methods,” *Foundations and Trends in Human-Computer Interaction*, vol. 11, no. 2, pp. 63–125, 2017.
2. C. Alexander, *The timeless way of building*, vol. 1. New York: Oxford University Press, 1979.
3. W. Edwards Deming, *Quality, productivity, and competitive position*. MIT Press, 1982. See https://en.wikipedia.org/wiki/Total_quality_management, https:

//en.wikipedia.org/wiki/Kaizen, and https://en.wikipedia.org/wiki/Eight_dimensions_of_quality.

4. J. J. Gibson, "The theory of affordances. the ecological approach to visual perception," 1979.
5. M. d'Inverno, M. Luck, P. Noriega, J. A. Rodriguez-Aguilar, and C. Sierra, "Communicating open systems," *Artificial Intelligence*, vol. 186, no. 0, pp. 38 – 94, 2012.
6. P. Noriega, H. Verhagen, M. d'Inverno, and J. Padget, "A manifesto for conscientious design of hybrid online social systems," in *Coordination, Organizations, Institutions, and Norms in Agent Systems XII Revised Selected Papers*, pp. 60–78, 2016.
7. High-Level Expert Group on AI (AI HLEG), "Ethics guidelines for trustworthy AI," Apr. 2019.
8. The IEEE Global Initiative on Ethics of Autonomous and Intelligent System, "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, first edition," 2019.
9. D. North, *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1991.
10. P. Noriega, J. Padget, H. Verhagen, and M. d'Inverno, "Towards a framework for socio-cognitive technical systems," in *Coordination, Organizations, Institutions, and Norms in Agent Systems X*, pp. 164–181, Berlin / Heidelberg: Springer, 2015.
11. P. Noriega, J. Padget, and H. Verhagen, "Anchoring online institutions," in *Anchoring Institutions. Democracy and Regulations in a Global and Semi-automated World* (P. Casanovas and J. J. Moreso, eds.), Springer, 2021. In press.
12. H. Aldewereld, O. Boissier, V. Dignum, P. Noriega, and J. Padget, eds., *Social Coordination Frameworks for Social Technical Systems*. Springer, 2016.
13. R. Christiaanse, A. K. Ghose, P. Noriega, and M. P. Singh, "Characterizing artificial socio-cognitive technical systems," 2014.
14. S. H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," in *Advances in experimental social psychology*, vol. 25, pp. 1–65, Elsevier, 1992.
15. I. van de Poel, "Embedding values in artificial intelligence (AI) systems," *Minds and Machines*, vol. 30, no. 3, pp. 385–409, 2020.
16. M. Johansson, H. Verhagen, and Y. Kou, "I am being watched by the tribunal: Trust and control in multiplayer online battle arena games," in *Proceedings of FDG 2015*, 2015.
17. S. Tomkinson and B. van den Ende, "'thank you for your compliance': Overwatch as a disciplinary system," *Games and Culture*, p. 15554120211026257, 2021.

Pablo Noriega is a tenured scientist in the Artificial Intelligence Research Institute of the Spanish National Research Council (IIIA-CSIC). His research interests are mostly in regulated multiagent systems and agreement technologies. Lately, he has been active in modelling value-driven policy making. Contact: pablo@iiia.csic.es

Harko Verhagen is a senior lecturer in Computer and Systems Sciences at Stockholm University. He has been working on combining social science theories and models with agent research, and applying these in social simulation as well computer game studies since the early 1990s. Contact: verhagen@dsv.su.se

Julian Padget is a reader (associate professor) in Artificial Intelligence at the University of Bath (UK). He has been working on norm representation and reasoning for multiagent systems since the mid 1990s, while exploring applications in virtual reality, generative narrative, policy-making and legal reasoning. Contact: j.a.padget@bath.ac.uk

Mark d'Inverno is a Professor of Computer Science at Goldsmiths, University of London. He was formerly Pro-Warden Research and Enterprise and subsequently Pro-Warden International. He has been PI and Co-I in a range of UK and internationally funded projects and has worked on the interface of AI, social science and Arts practice for several decades. Contact: dinverno@gold.ac.uk